Data wrangling report

Introduction

This report aims to document the data wrangling efforts carried out in the data wrangling project of the ALX-T Data Analyst Nanodegree. The wrangling efforts were conducted in four processes i.e. data gathering, data assessing, data cleaning and data storage.

Data gathering

To begin with, I manually downloaded the WeRateDogs Twitter archive file named ‘twitter_archive_enhanced.csv’. This csv file is the tweet archive of twitter user WeRateDogs and was provided by the Udacity Team. Subsequently, I programmatically downloaded the image_predictions tsv file which was hosted on Udacity’s servers using the Requests Library. Lastly, I queried the Twitter API using the Tweepy library to gather additional data such as retweet count and favorite count.

Data assessing

All the gathered data was assessed both visually and programmatically and an assessment report was produced which highlighted all the quality and tidiness issues identified in the assessment. The visual assessment was carried out in both the Jupyter Notebook and in Microsoft Excel while the programmatical assessment was carried out in the notebook. In total, there were 12 quality issues and 5 tidiness issues identified in three pieces of gathered data. Examples of identified quality issues include invalid dog names such as ‘a’, ‘the’ etc. and incorrect dog ratings that mainly stemmed from combined rating of multiple dogs in one picture. Tidiness issues example include 4 different columns for dog stage.

Data cleaning

Following the assessment, all the 12 quality issues and the 5 tidiness issues were successfully cleaned. The cleaning process religiously followed the define-code-test framework. The order of the framework is that you briefly define what you intend to do, write the piece of code to achieve the stated objective then thoroughly test to ensure your code successfully achieved the declared objective. After all the issues were cleaned, a master dataset was produced named ‘twitter_archive_master’. This dataset was a result of merging retweet count, favorite count, dog breed prediction and confidence columns from the predictions file and json file into the twitter archive file.

Data storage

The final stage of the data wrangling phase consisted of the storage of the newly created master dataset. The dataset was successfully stored in a csv file.